Abstract Title Page

Title: Designing a sample selection plan to improve generalizations from two scale-up experiments

Author(s):
Elizabeth Tipton, Kate Sullivan, Larry Hedges, Michael Vaden-Kiernan, Geoffrey Borman & Sarah Caverly

Abstract Body.

Background / Context:
In 2009, IES funded a scale-up, cluster randomized trial evaluating the Open Court Reading curriculum (OCR) based on extensive use across its 40 year history. OCR is a core reading program for elementary school students. The curriculum is built on research-based practices cited in the National Reading Panel (National Reading Panel, 2000) report and emphasizes phonemic awareness, phonics, fluency, vocabulary, and text comprehension. In 2010, IES funded a scale-up, cluster randomized trial evaluating the Everyday Mathematics (EM) curriculum after almost three decades of research and development, the widespread use of the program, and promising evidence of program efficacy (Slavin & Lake, 2007; What Works Clearinghouse, 2007, 2010). EM is a comprehensive, reform-based mathematics curriculum for elementary school students.

In an effort to maximize sample size and statistical power, these two trials were combined in terms of design, recruitment, sampling frame, and study samples. The design used is a three-level randomized block design, in which districts are recruited into the study and then within these districts 4 schools are recruited. Two of these schools are randomly assigned to receive OCR and two of the schools are assigned EM. The opposing schools act as the statistical controls. Based on a power analysis, it was determined that 15 districts would be recruited into the combined scale-up study, for a total of 60 schools. In order to address face validity concerns, the initial plan was to recruit districts by blocking on two variables: Census district and geographic locale.

While this recruitment plan addressed generalizability for two variables, it did not provide a clear strategy for selecting districts when more than two variables were of interest. In order to address this problem, we turned to recent work that focuses on improving generalizations from experiments to particular populations in the retrospective case through a new application of propensity score matching methods (Tipton, 2011; Hedges & O'Muircheartaigh, 2011; Stuart, Cole, Bradshaw, & Leaf, *in press*; Roschelle, Tatar, Hedges, & Tipton, 2010). In this paper we extend this work to the prospective sampling case, and use propensity score matching to develop a strategic recruitment plan that leads to a sample that is representative of the population of interest. This is particularly important here since the goal for both the OCR and EM experiments is to evaluate the programs at scale, which requires both well-defined populations and sample selection procedures.

Purpose / Objective / Research Question / Focus of Study:
The purpose of this paper is to introduce a strategic sample selection method and to report on the use of this method in the OCR and EM sample recruitment case introduced above. An important step in the sample selection process is the definition of the population and sampling frames, as well as variables for matching. In this section, we discuss the particulars of these choices in the OCR and EM cases.

The first step in this method is to define the populations of interest. We decided to define the EM population as the set of school districts like those currently using the EM curriculum and the OCR population as the set of school districts that are like those currently using the OCR curriculum. We chose to focus on current users since this population is well defined, whereas the

population of future users is not. Through our partnership with McGraw-Hill (MGH), we gained access to data describing district level sales figures from 2008-2010. During this three-year period, 4,888 school districts from all 50 states and the District of Columbia purchased the OCR and/or EM materials from MGH. Of these, 1,472 purchased only the OCR curriculum, 2,743 purchased only the EM curriculum, and 673 purchased both.

The second step in this method is to determine the sampling frame, which is the set of units that is eligible for inclusion in the experiment. The generalization population and the sampling frame may be the same, but in many cases are not; for example these two groups may differ if including certain population units in the experiment would be cost prohibitive (based on financial, geographic, or political concerns) or would jeopardize the internal validity of the experiment. This second case occurs here, where including current users would be impossible since there would be no relevant comparison group. As a result, the sampling frame for this study was determined based on two criteria: 1) sales history and 2) school level eligibility. For sales history, districts passed the first eligibility criteria if they had no sales history in the previous three years (2008 – 2010). Districts were then evaluated based on the availability of schools meeting sampling requirements set forth in power estimates, stipulating that districts include at least four elementary schools with at least 44 students in each of grades Kindergarten through fifth grade. Based on these criteria, the eligible population of school districts for both the OCR and EM studies included 675 school districts across the country.

The goal of this paper is to develop and implement a method for selecting 15 school districts out of the 675 districts that are eligible for inclusion and that best represent the OCR and EM user populations. Since a random sample is infeasible, we instead focus on choosing a sample that has a similar composition to that of the generalization populations on a set of key covariates. These are the covariates that may explain heterogeneity in the district average treatment effects, since the population average treatment effect depends on the composition of the population with regard to these covariates. To this end, we decided to match the compositions of the sample and populations on 11 variables from the Common Core of Data (CCD). These variables are listed in Table 1, and include region, urbanicity, racial composition, measures of poverty and education levels in the districts.

INSERT TABLE 1 HERE

Significance / Novelty of study:
Currently scale-up study samples are selected to be representative by blocking on a small number of important variables. When the number of districts or schools included in a sample is to be small and the number of covariates is large, however, the blocking approach will lead to empty blocks. For example, with only 15 districts in an experiment, using more than 3 dichotomous variables leads to empty blocks. In the face of this, current practice is often to limit the number of variables for matching to a very small number (e.g. 1 or 2). The propensity score matching approach we develop here allows for matching on a much larger set of variables, including both continuous and categorical variables. By matching the sample and population compositions on a larger set of covariates, sample selection bias in the estimate of the population average treatment effect can be reduced or eliminated.

A key feature of the method developed here is that it is practical and flexible. Clearly a random sample would be ideal; however, in the case in which the response-to-recruitment-rate is small (as often occurs when recruiting for experiments) many of these benefits are diminished. By using propensity score matching methods, we are able to target eligible units for recruitment so that the overall sample and populations are balanced on these key variables. When these targeted units will not agree to be in the experiment, our method provides a ranked list of similar units for recruitment. In some cases, there are many possible eligible districts to recruit from. In other cases, there are not as many choices. The extra benefit of this approach is that it helps researchers determine which types of districts may be the most difficult to recruit (i.e. those with very few similar eligible units), which helps in the allocation of resources during the recruitment process.

Statistical, Measurement, or Econometric Model:
The method we develop and implement here works in both the single generalization population case and the more complex two-generalization populations case. The problem particular to the OCR and EM study is that one sample of 15 districts had to be selected for two separate generalization populations: (1) the OCR user population and (2) the EM user population. In this section we give a brief overview of the method we developed for selecting districts for this dual purpose. The approach we develop here is a stratified sampling scheme, which improves balance and reduces bias (Cochran, 1968; Groves et al, 2009).

(1) Define the generalization populations and sample frame. For each of the two populations, combine the population and eligible units into a single data set and estimate propensity scores based on a set of covariates that may explain variation in district average treatment effects. Here the propensity score is the probability of being in the set of eligible units ($Z=1$) conditional on a set of covariates that explains variability in district-average treatment effects,

$$e_p(x) = Pr(Z=1|X=x, P=p),$$

where P is the population of interest ($p=1,2$). These propensity scores can be estimated using a logistic regression model.

(2) For each of the two populations, stratify the estimated propensity scores so that each stratum contains $1/m^{th}$ of the population, where m is a divisor of the total sample size n. The sample should be allocated to the strata using proportional allocation, so that n/m of the sample cases are allocated to each of the m strata. For example, if $n=15$ districts and $m=3$ strata are used, then $15/3=5$ districts will need to be recruited from each stratum.

(3) When there are two populations, plot the bivariate distribution of the propensity scores for the units in the sampling frame only (i.e. the eligible units). In order to determine how many units should be sampled from each of the $m*m$ combined strata, use iterative proportional fitting (with rounding) based on the number of eligible units in each stratum and the population specific marginal sample sizes.

(4) For each of the $j=1...m*m$ strata calculate the sampling fraction. This is the ratio of sampled cases to eligible units, n_j/N_{ej}. Strata with larger sampling fractions will require greater resources in the recruitment process; recruitment should therefore start with these difficult strata and end with the stratum with the smallest sampling fraction.

(5) Within each of the $j=1...m*m$ strata, for each unit calculate the distance,

$$D_{ij} = (logit(e_1(x_{ij})) - M_j(logit(e_1(x_{ij})))^2 + (logit(e_2(x_{ij})) - M_j(logit(e_2(x_{ij})))^2,$$

where $M_j(logit(e_p(x))$ is the average value of the logit values for units in the population in stratum j and for populations $p = 1, 2$. Within each stratum, rank the units from smallest to largest D_{ij} value. In recruitment, units with smaller ranks will be preferred, since these units are closer to the stratum averages, which ensures greater overall balance in the covariates for the combined sample and populations.

Usefulness / Applicability of Method:
We developed this method in relation to the problem of selecting the sample for the OCR and EM combined study. For each of the two populations, we estimated propensity scores using the **matchit** package in **R** (Ho, Imai, King, & Stuart, 2007). Figure 1 shows the bivariate distribution of propensity scores for the eligible units and includes lines signifying the 9 strata created by using $m=3$ strata in both the OCR and EM populations. Note that these propensity scores are positively correlated, suggesting that the OCR and EM populations are more similar than different. Within each of the $m*m=9$ strata, the distance measure D_{ij} defined above was calculated and units were then ranked from smallest to largest.

INSERT FIGURE 1 HERE

Since $n=15$ and $m=3$, each of the OCR and EM marginal strata were allocated $15/3=5$ districts for recruitment into the study. Using iterative proportional fitting, these cases were allocated to the 9 bivariate strata. Table 2 shows the number of districts that must be recruited relative to the number of eligible units for each of these 9 bivariate-strata. Note that one stratum does not contain any eligible units. These strata are ranked from those with the largest sampling fractions (the districts more difficult to recruit) to the smallest (column "Stratum sampling order").

INSERT TABLE 2 HERE

Based on this analysis the recruitment team was given three instructions. First, sample from the most difficult strata first and give a greater proportion of resources to recruitment efforts in these strata. Second, within each stratum, use the distance ranks to determine the order in which districts are contacted or considered for recruitment. Third, when the last stratum (here Stratum 11) is reached, recalculate distances to offset any residual imbalances that arise from sub-optimal recruitment in the other strata.

In the final paper, we will present a discussion of how this recruitment plan fared in practice, what issues arose, and how the composition of the final recruited sample compared to the two populations of interest.

Conclusions:
In this paper we present a new method for sample selection for scale-up experiments. This method uses propensity score matching methods to create a sample that is similar in composition to a well-defined generalization population. The method we present is flexible and practical in the sense that it identifies units to be targeted for recruitment, and when they are not available, identifies similar units for replacement. Additionally, this method helps researchers determine which areas of the population may be most difficult to recruit from, enabling resources to be allocated accordingly.

Appendices

Appendix A. References

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295-313.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., & Singer, E. (2009). *Survey methodology* (p. 461). John Wiley and Sons.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007) Matchit: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software. Software, http://gking.harvard.edu/matchit/.

Roschelle, J., Tatar, D., Hedges, L.V., & Tipton, E. (2010) "Two perspectives on the generalizability of lessons from scaling up SimCalc." Paper presented at the Annual Conference for the Society for Research Synthesis Methods. Washington, DC.

Hedges, L.V. & O'Muircheartaigh, C.A. (*under review*) Improving generalization from designed experiments. *Working Paper.*

National Reading Panel. (2000). Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Washington, DC: Natrional Institute of Child Health and Human Development.

Slavin, R. E., & Lake, C. (2007). Effective programs in elementary mathematics: A best-evidence synthesis. Baltimore, MD: Johns Hopkins University.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (in press). The use of propensity scores to assess the generalizability of results from randomized trials. *Forthcoming in The Journal of the Royal Statistical Society, Series A. PMC Journal.*

Tipton, E. (2011). Improving the external validity of randomized experiments using propensity score subclassification. *Working Paper.*

What Works Clearinghouse. (2007). Elementary School Math. Retrieved April 18, 2011, from http://ies.ed.gov/ncee/wwc/reports/elementary_math/topic/references.asp

What Works Clearinghouse. (2010). Intervention: Everyday Mathematics. Retrieved April 18, 2011, from http://ies.ed.gov/ncee/wwc/reports/elementary_math/eday_math/

Appendix B. Tables and Figures

Table 1. Characteristics of inference and eligible population

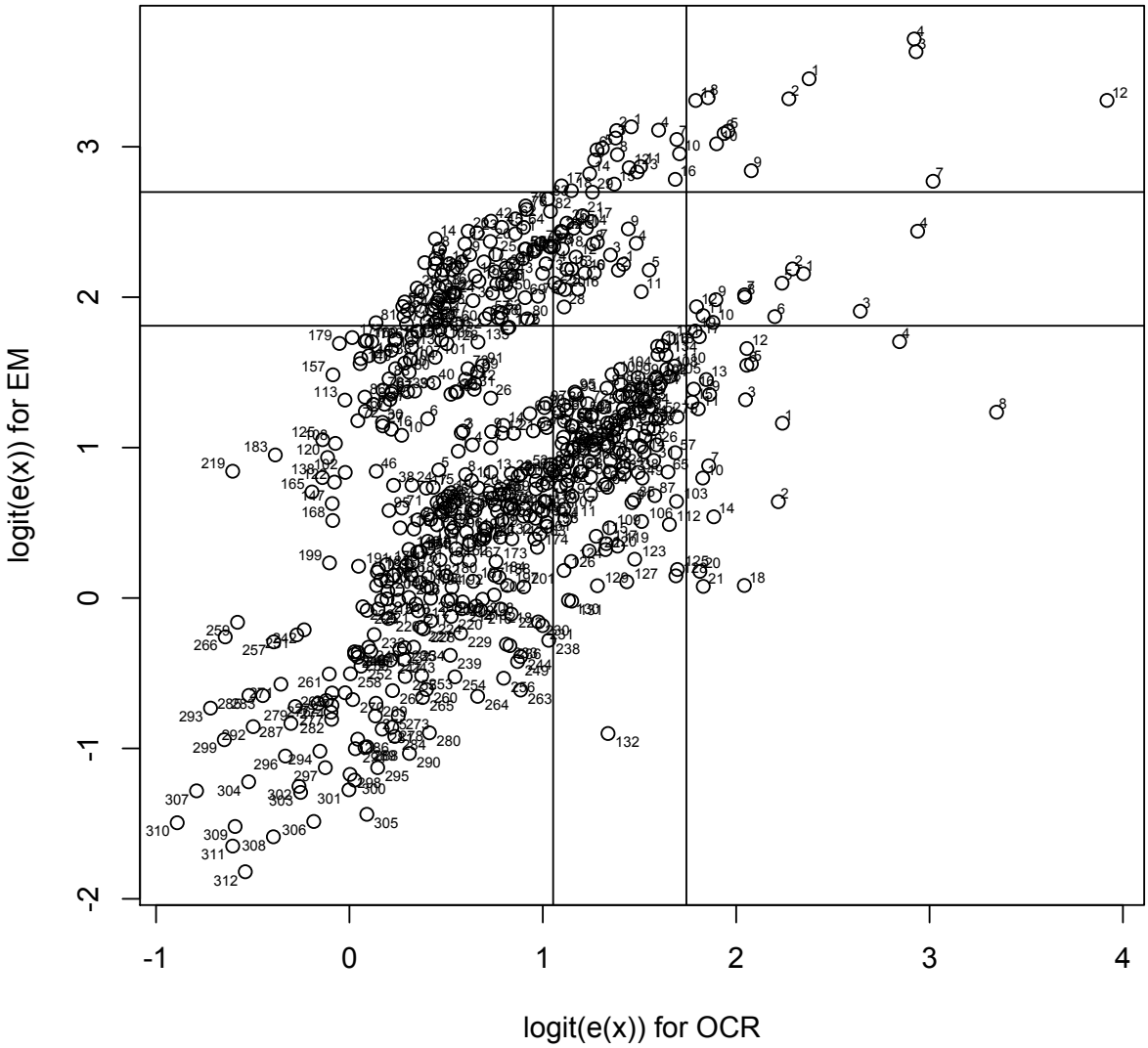
	Likely adopters N = 675				Typical Users - EM N = 3478				Typical Users - OCR N = 2173			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Community Characteristics												
Educational Attainment												
% 8th grade or less	9.9%	7.6%	0.4%	57.5%	7.7%	5.9%	0.0%	63.0%	11.4%	9.5%	0.0%	63.2%
% less HS grad	15.5%	6.0%	1.6%	31.6%	13.5%	6.0%	0.2%	36.6%	16.3%	6.2%	0.9%	50.0%
% HS grad	36.8%	9.1%	6.3%	58.1%	40.9%	11.8%	3.1%	69.4%	39.0%	10.5%	4.4%	67.3%
% post secondary	37.9%	16.4%	5.8%	91.0%	37.8%	18.4%	5.5%	96.1%	33.3%	17.0%	1.8%	93.6%
Financial stats: Census area												
% in labor force	64.6%	6.5%	34.7%	81.8%	64.9%	7.1%	21.4%	86.1%	62.0%	7.8%	17.6%	91.6%
Median income overall	54046.2	18114.2	16411.0	173777.0	54452.0	21061.4	17083.0	200001.0	47940.0	19042.5	17061.0	200001.0
Percent of 5-17 year olds in poverty	13.6%	9.1%	1.1%	60.7%	11.9%	9.2%	0.0%	67.8%	16.7%	11.0%	0.0%	70.4%
School District Characteristics												
Urbanicity of districts												
% Urban area	24.3%	42.9%	0.0%	100.0%	7.8%	26.9%	0.0%	100.0%	9.2%	28.8%	0.0%	100.0%
% Rural area	13.9%	34.6%	0.0%	100.0%	23.2%	42.2%	0.0%	100.0%	26.0%	43.9%	0.0%	100.0%
% Suburb	34.7%	47.6%	0.0%	100.0%	29.2%	45.5%	0.0%	100.0%	20.8%	40.6%	0.0%	100.0%
% Town	13.2%	33.9%	0.0%	100.0%	19.0%	39.2%	0.0%	100.0%	18.9%	39.1%	0.0%	100.0%
Geographic location of districts												
% Northeast	21.2%	40.9%	0.0%	100.0%	27.5%	44.7%	0.0%	100.0%	16.1%	36.8%	0.0%	100.0%
% Midwest	17.3%	37.9%	0.0%	100.0%	40.5%	49.1%	0.0%	100.0%	16.2%	36.8%	0.0%	100.0%
% South	30.8%	46.2%	0.0%	100.0%	15.8%	36.4%	0.0%	100.0%	35.5%	47.9%	0.0%	100.0%
% West	30.7%	46.1%	0.0%	100.0%	16.2%	36.8%	0.0%	100.0%	32.3%	46.8%	0.0%	100.0%
District expenditures per student	11691.4	3692.1	5996.8	37790.3	13506.2	8670.7	500.5	188527.3	13529.6	10638.4	500.5	177613.1
Student Characteristics												
Averge number of students in district	12322.9	11650.6	1760.0	100685.0	6605.4	21857.9	0.0	684143.0	8596.4	26324.1	0.0	684143.0
Race/ethnicity of district												
Percent White	52.2%	28.7%	0.1%	98.6%	73.4%	26.5%	0.0%	100.0%	58.2%	31.2%	0.0%	100.0%
Percent Black/African American	22.9%	24.8%	0.1%	99.9%	9.6%	15.6%	0.0%	95.0%	19.5%	25.2%	0.0%	100.0%
Percent Hispanic	1.1%	4.3%	0.0%	80.1%	2.2%	9.8%	0.0%	99.8%	3.7%	13.4%	0.0%	100.0%
Percent other (Asian, Pacific Islander, American Indian, Native Alaskan, 2 or more races)	7.0%	9.8%	0.0%	80.8%	5.3%	11.1%	0.0%	100.0%	7.2%	14.7%	0.0%	100.0%
Percent of students identified as ELL	10.3%	12.2%	0.0%	69.6%	4.3%	8.1%	0.0%	88.4%	8.7%	13.0%	0.0%	88.4%
Percent of students who receive free & reduced price lunch	44.2%	22.2%	0.0%	99.4%	36.0%	21.9%	0.0%	99.5%	43.8%	23.2%	0.0%	99.5%

Table 2. Nine strata for recruitment and allocation of sample and resources

Stratum	Stratum sampling order	# Eligible districts	# Districts in sample	Sampling fraction	Proportion of resources
31	0	0	0	0.000	0%
33	1	12	3	0.250	42%
32	2	18	2	0.111	18%
23	3	12	1	0.083	14%
22	4	29	2	0.069	11%
13	5	21	1	0.048	8%
21	6	83	2	0.024	4%
12	7	132	1	0.008	1%
11	8	312	3	0.010	2%

**Note: Proportion of resources it the sample fraction for a particular stratum as a fraction of the total sum of the sample fractions across strata.*

Figure 1: Bivariate distribution of estimated propensity score logits



**Note: Only eligible districts are included in the bivariate plot. These are labeled by the within stratum ranks, where lower ranked districts are preferred.*